LESSON 3 – BASIC STATISTICAL APPLICATIONS

Dr. Jack Hong
Adjunct Faculty, Lee Kong Chian School of Business, SMU
Co-founder, Research Room Pte. Ltd.

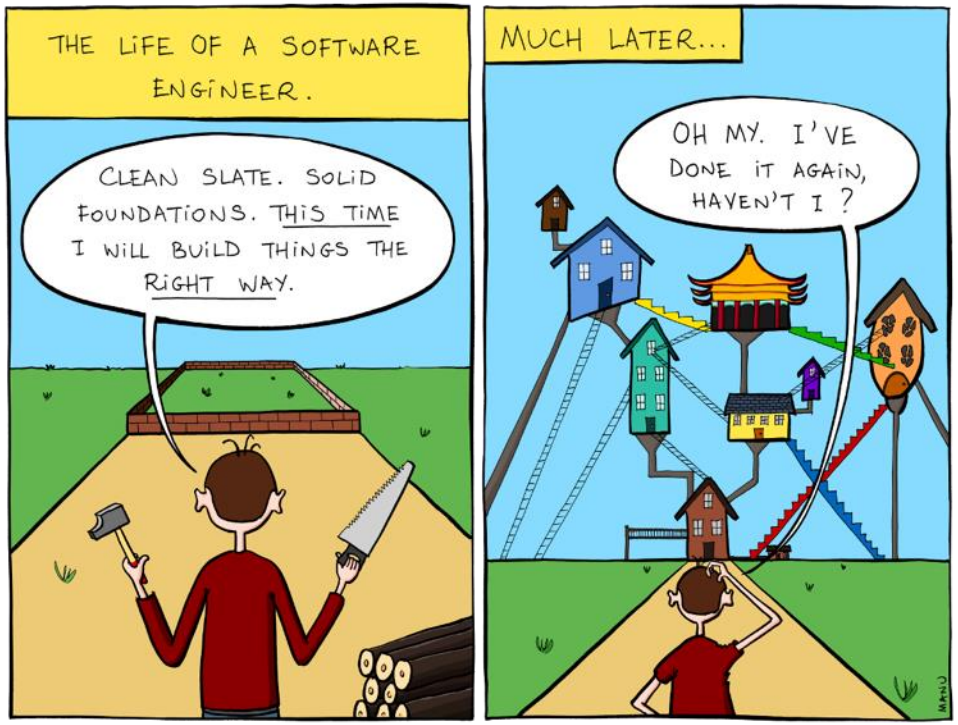
BEFORE WE START…

- What's the difference between statistical and machine learning models?
  - They are more similar than most people realize
  - Both models attempt to extract patterns from data that estimates relationships between variables
- Statistics focus on describing the data and validating hypotheses about the data (inference)
  - Inference: Draw conclusions about a set of information, guided by what the researcher knows/hypothesizes
    - **Usually backed by theories, even if model is stylized**
    - Exploratory: Discover new features
    - Confirmatory: Prove hypotheses to be True or False
- Machine learning focus on extracting patterns (mining) from data and using it for prediction
  - Data mining starts from "I have no idea what I'm looking for"
    - Finding patterns and anomalies to form insights
  - Statistical models that extract patterns from data are found in machine learning as well (e.g. regressions)

ESTIMATING RELATIONSHIPS

BEFORE YOU START PLAYING WITH THE DATA:

VISUALIZE THE LAYOUT!



WHAT KIND OF VARIABLES ARE THERE?

- Variable (Random Variable)
  - A characteristic that takes different values for different entities (e.g. person, stock)

Person	Date	Gender	AgeRange	Citizenship	MaritalStatus
S1234567A	2014-10-31 20:41:49	Male	11	2	2
S1234568B	2014-11-01 10:40:33	Female	3	2	1

  - 2 types:
    - Quantitative
    - Categorical
      - Often used for qualitative data

Variable	Quantitative	Categorical
Age	Year-month-day-hour-min-secs	Gen X; Gen Y; Millennials
Salary	$ and cents	Grade A, B, C Supergrade

CATEGORICAL VARIABLES

- 3 subtypes of Categorical Variables
  - Nominal: Unordered
  - Dummy: Takes only 2 values (1 or 0), typically used to represent (Yes, No) or (Success, Failure)
  - Ordinal: Ordered but no defined distance between intervals
- The first 2 subtypes are the most common

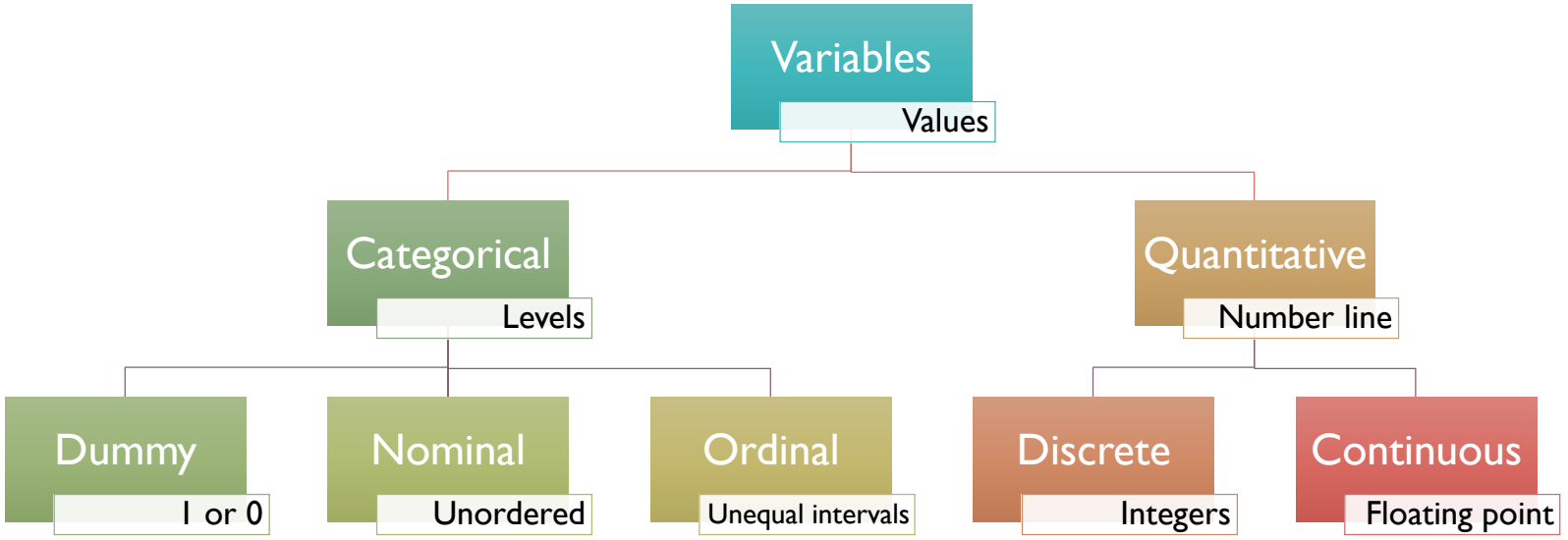
Variable	Quantitative	Categorical		
		Nominal	Ordinal	Dummy
Education	No. of Years	Law; Finance; Physics; Psychology; Business; Economics	PSLE; O-Levels; A-Levels; Diploma; Undergrad; Postgrad	Have Bachelor Degree?
Notes:	Can rank in terms of duration. Difference between ranks are uniform.	Unable to rank.	Can rank in terms of academic potential, but the difference between ranks are not uniform.	Yes, No

QUANTITATIVE VARIABLES

- 2 subtypes of Quantitative Variables
  - Distance of intervals are uniform
    - The effect of an increment from 2 to 3 is of the same magnitude as an increment from 3 to 4
  - Discrete: Integers (e.g. 0, 1, 2, 3, 4)
  - Continuous: Infinite possibilities within a range (i.e. floating numbers like 3.14159265359 where the number of decimals can go to infinity)
  - Both discrete and continuous variables are treated in the same manner in statistical work

IMPORTANT NOTE ABOUT VARIABLES

- Model designs and interpretations are distinctive between
  - quantitative and categorical variables
  - nominal and ordinal variables
- Example: Salary increases with Age (quantitative, ordinal)
  - As age increases, we tend to see salary increasing as well
- Example: Salary increment in 2018 is positive with Race A, B, and negative with Race C (categorical, nominal)
  - Races do not have an ordinal scale (e.g. There is no logic why Race A is greater or smaller than C)



DATA SETUP

- **Cross-sectional**
  - Each observation is an entity (individual, firm etc) with associated information at a specific point in time
  - Various characteristics that are observed in conjunction with the observation are known as features
  - Data analytics is generally about associating characteristics with an outcome (e.g. Earnings before interests and taxes - EBIT), then using the strength or probability of the association to predict future EBIT given a set of features
    - Example: Given the strength of the associations of the characteristics of a company and its EBIT in 2010, we can predict 2011 EBIT using the characteristics observed in 2011.

Features

Observations

Company Name	Exchange:Ticker	Excel Company ID	Year	Total Revenue	Gross Profit	EBITDA	EBIT
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2006	50.2	12.6	4.76	3.56
3SBio Inc. (SEHK:1530)	SEHK:1530	IQ24164591	2006	64.5	14	11.7	11.4
800 Super Holdings Limited (Catalist:5TG)	Catalist:5TG	IQ137292017	2006	19.1	3.45	4.24	2.22

DATA SETUP

- **Time Series**
  - Dataset consist of 1 entity, tracked over time
    - Sample is not random, need to consider corrections on many fronts
  - Important for trending and seasonality studies
  - Important for within-person studies, where data is collected on a person over time (**"The best predictor of future behavior is past behavior."**)
  - Note: Features that does not vary between observations (such as Exchange:Ticker and Excel Company ID in the following table) are not useful for algorithms

Features

Observations

Company Name	Exchange:Ticker	Excel Company ID	Year	Total Revenue	Gross Profit	EBITDA	EBIT
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2006	50.2	12.6	4.76	3.56
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2007	44.7	13.3	5.62	4.61
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2008	57.9	16.8	5.39	3.75
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2009	114.1	39	23.9	22.7
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2010	97.8	33.2	18.8	17.5
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2011	92.3	20.5	5.02	3.51
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2012	80.9	19	6.4	5.22
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2013	89.3	19.9	6.74	5.42
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2014	120.2	24.6	9.38	7.73

DATA SETUP

- Panel data (we encounter such datasets most of the time)
  - Pooled cross-sectional and time-series data
  - We treat this similarly to a normal cross-sectional dataset
  - Accounting for time differences can be easily accomplished using time categorical variables, and can be extended to any subset of groupings
  - Useful for tracking the same entities over time

Company Name	Exchange:Ticker	Excel Company ID	Year	Total Revenue	Gross Profit	EBITDA	EBIT
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2006	50.2	12.6	4.76	3.56
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2007	44.7	13.3	5.62	4.61
3Cnergy Limited (Catalist:502)	Catalist:502	IQ79611391	2008	57.9	16.8	5.39	3.75
3SBio Inc. (SEHK:1530)	SEHK:1530	IQ24164591	2006	64.5	14	11.7	11.4
3SBio Inc. (SEHK:1530)	SEHK:1530	IQ24164591	2007	346.9	11.4	-0.287	-1.42
3SBio Inc. (SEHK:1530)	SEHK:1530	IQ24164591	2008	366	10.1	-2.66	-3.85
800 Super Holdings Limited (Catalist:5TG)	Catalist:5TG	IQ137292017	2006	19.1	3.45	4.24	2.22
800 Super Holdings Limited (Catalist:5TG)	Catalist:5TG	IQ137292017	2007	22.2	3.06	3.61	2.1
800 Super Holdings Limited (Catalist:5TG)	Catalist:5TG	IQ137292017	2008	17.3	0.958	1.76	0.172

$$Return\ on\ Assets = R\&D + CEO\ pay + i.Year + i.Company + i.CEO$$

Dependent Variable

Independent variables

Categorical effects: Year, Company and Person dummies

IN THE NEXT SECTION…

- We will cover 2 basic but very important statistical models:
  - Linear regression
  - Logistic regression

- Most studies involving statistical modeling rely on these 2 models
  - Linear regression estimates the strength of the relation between X and Y
  - Logistic regression estimates the increased/decreased probability of observing Y when X is present
  - There are many advanced variants but we will not introduce them in this pre-course
    - Email me for resources if you are interested to learn more

STATISTICAL MODEL TECHNIQUES

THE LINEAR REGRESSION

REGRESSION ANALYSIS

- The preferred tool most relied upon by scientists
- A way to quantify relationships in numerical form
  - More flexible and awesome than what your undergraduate instructor led you to believe!
  - Creativity required:
    - Collecting/ creating the right measures
    - Finding the right setup (explanatory variables that matter)
    - Quantifying the qualitative (sentiments, opinions etc.)

LHS
Dependent variable
Response variable
Explained variable
Regressand

$$y = f(X), where\ X \in x_1, x_2, x_3, \ldots$$

RHS
Independent variable
Explanatory variable
Regressor
Covariate
Control variable

REGRESSION ANALYSIS

$$y = f(X), where\ X \in x_1, x_2, x_3, \ldots$$

- y is known as the outcome variable
  - In machine learning speak: This is the variable that we want to predict using a range of characteristics ($x_1, x_2, x_3, \ldots$)
  - In statistical speak: This is the variable that we want to explain using a range of characteristics ($x_1, x_2, x_3, \ldots$)
- X is the notation for a collection of ($x_1, x_2, x_3, \ldots$)
  - In machine learning speak: These are the features that we want to use to predict y
  - In statistical speak: These are the explanatory variables that we want to use to explain how y moves
- f() is simply a notation for function
  - Tells us that X are connected via some operators without specifying what they are
  - It could be $x_1 + x_2 + x_3$ or $x_1/(x_2 + x_3)$ or something really complex

THE LINEAR REGRESSION

- If f() is linear, then the equation can be written as a linear combination of X:

$$y = x_1 + x_2 + x_3 + \cdots$$

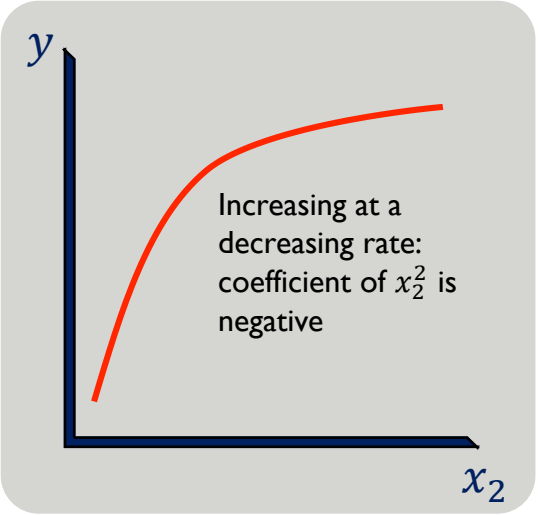
- Linear equations are not as linear as it seems. By creating variables such as squared terms (age x age) or interaction terms (age x qualification), we introduce non-linearity to the regression model

$$-\ln(y) = x_1 + x_2 + x_2^2 + x_3 + x_1 . x_3 + \cdots$$

Loglinear transformation, interpret as % change in y

x_2-squared, captures 2nd derivative (d^2y/dx^2) – rate of change

Interaction term, similar to squared terms, captures rate of change between 2 variables



THE LINEAR REGRESSION

- The key purpose of the linear regression is estimate the strength of the relationship between y and X
- The estimation model takes the following form

$$\hat{y} = \alpha + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

  - where α is called the intercept, and gives the average value of y
  - $\hat{\beta}$ are known as coefficients, and each x has a coefficient that indicates the strength of its relationship with y
- In regression analysis, relationship strength is measured by the co-movement of y and x
  - Looking at the entire dataset, when x moves by 1 unit, how much does y move on average?
    - If on average, whenever we observe a change of 1 unit in the value of x_1, we observe a change of 0.5 in the value of y -> the coefficient $\widehat{\beta_1}$ is 0.5

SIMPLE EXAMPLE

- Upload "statistical models – linear regression.ipynb" into your Jupyter dashboard
- In this example, given the length of eruption time, we want to estimate the amount of waiting time until the next eruption
- The linear regression equation is
  - $\text{waiting\_}time = \alpha + \beta_1 eruption$
  - where α is an estimation of the average waiting time (average of the LHS variable)
  - and β_1 is a coefficient that measures the relationship between length of eruption time and waiting time until next eruption (relationship between the LHS variable and the target RHS variable)

RESULTS OUTPUT FROM R

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  33.4744     1.1549   28.98   <2e-16 ***
eruptions    10.7296     0.3148   34.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.914 on 270 degrees of freedom
Multiple R-squared:  0.8115,    Adjusted R-squared:  0.8108
F-statistic:  1162 on 1 and 270 DF,  p-value: < 2.2e-16
```

- Results
  - α is 33.47 min, implying that the average waiting time to next eruption is approximately 33 min
  - β_1 is 10.73, implying that a 1 min eruption length of time increases the waiting time to next eruption by approximately 11 min

THE LINEAR REGRESSION (T-VALUE)

- However, knowing the coefficient is not sufficient
- In statistics, we have tests to show whether the coefficient is significant or not
  - This test of significance determines whether the value of the coefficient is statistically different from 0 or not
  - If the coefficient is 0, this implies that there is no relationship
  - The test for this significance involves dividing the estimated coefficient by the standard error (the difference between all observed and predicted values)
    - If the standard error is small, then the coefficient is likely to be different from 0
    - The divided value is known as the t-value, and the cutoffs are at confidence levels (CI) of 90% (~ t-value above 1.6), 95% (~ t-value above 1.8), and 99% (t-value above ~1.97)
    - Most statistical packages use *, **, *** to denote 90%, 95%, and 99% CI
- In the results (faithful), the t-values for eruptions is 34.09, with ***.
  - Generally, any confidence above 90% is considered statistically significant
  - However, many industry only consider confidence above 95% as statistically significant

THE LINEAR REGRESSION (R^2 AND F-STATISTIC)

- There are 2 other important metrics in the results that determine whether our regression model is good or not
  - By good, we mean whether the explanatory variables (X) sufficiently explain what is happening to outcome (y)
- The first metric is the Adjusted R^2
  - This metric is obtained by comparing the variations in y against the variations in x
  - If there is a big overlap in both variations, it means that there is goodness of fit
  - In the results, adjusted R^2 is 0.8108 (81.01%), which implies that 81% of the variations in the data can be explained by the linear relationship (and this is a good fit)
- The second metric is the F-statistic
  - This metric compares whether the fitted model is any different from a model without any X (i.e. just taking the average value of y)
  - In other words, F-statistic show whether R^2 is significantly different from 0 or not
  - In the results, p-value for the F-statistic is very small (< 0.01). This implies that the fitted model is better than random chance

PREDICTING WITH LINEAR REGRESSION

- With the fitted model, we can predict the waiting time until the next eruption, given the length of the current eruption
- Note that statistical models are generally good for inference
  - Inference is what we have done in the previous few slides
    - Whether an explanatory variable is statistically significant
    - Whether the model has a good fit to historical data
- We generally rely on machine learning (ML) models for prediction
  - ML models have an uncanny ability to fit mathematical models to data, often in a black box approach
  - ML tests are not that concerned about inferences, but rather how accurate the ML models predict future observations

MULTIVARIATE REGRESSIONS AND CATEGORICAL VARIABLES

- What happens when we have multiple variables on the RHS? How do we interpret it?

$$\hat{y} = \alpha + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3$$

- Interpreting coefficients in a multivariate regression:
  - Holding x_2 and x_3 fixed, the relationship between y and x_1 is given by $\hat{\beta}_1$
  - Holding x_1 and x_3 fixed, the relationship between y and x_2 is given by $\hat{\beta}_2$
  - ceteris paribus approach

- "Stripping out" interpretation
  - $\hat{\beta}_1$ is the effect that remains after the effect of x_2 on y has been stripped out

MULTIVARIATE REGRESSIONS AND CATEGORICAL VARIABLES

- Categorical variables cannot be explained the same way as numerical variables
  - E.g. salary = age + age^2 + gender
- If we observe a positive coefficient for age
  - This means that salary increases the older you are
- If we observe a negative coefficient for age^2
  - This means that the salary is increasing at a decreasing rate the older you are
- If we encode gender as male = 0, female =1, transgender = 2
  - And we observe a positive coefficient for gender
  - We might want to conclude that transgender earns more than female, who earns more than male
  - This interpretation is wrong because there is no rationale why genders are ranked in an ordinal manner (higher number for age = older, but higher number for gender = ?)

MULTIVARIATE REGRESSIONS AND CATEGORICAL VARIABLES

- As such, categorical variables are interpreted via benchmarking against a base category
  - The equation actually looks like this:

    $salary = age + age^2 + female + transgender$

    - where each observation can only have 1 gender (i.e. if male, then female = 0, transgender= 0)
    - Where is male=1? Because male can be denoted by female=0 and transgender =0, therefore having another column for male = 1 is redundant

Person ID	salary	age	age x age	female	transgender
1	90,000	32	1024	1	0
2	184,500	28	784	0	1
3	48,650	42	1764	0	0

MULTIVARIATE REGRESSIONS AND CATEGORICAL VARIABLES

Person ID	salary	age	age x age	female	transgender
1	90,000	32	1024	1	0
2	184,500	28	784	0	1
3	48,650	42	1764	0	0

$$salary = age + age^2 + female + transgender$$

- As we denote male as female = 0 and transgender = 0, male becomes the benchmark category
  - The coefficient for female denotes how much more (or less) salary the female category earns against the male category
  - The coefficient for transgender denotes how much more (or less) salary the transgender category earns against the male category

EXERCISE

- Go through the codes in "statistical models – linear regression.ipynb" on the seatbelts dataset
- Load the titanic dataset and explore the data using the techniques demonstrated
  - Descriptive statistics
  - Inspecting, slicing data
  - Test of variances
  - Scatterplot
  - Linear regression (make sure you can interpret numerical and categorical variables)

STATISTICAL MODEL TECHNIQUES

THE LOGISTICS REGRESSION

LOGISTIC REGRESSION I

- We have mainly covered linear regressions with numerical response variable

- What if the response variable is binary?
  - {Success, Failure}, {Yes, No}
  - We can't use linear regression in this case because the model often predicts a value outside the range of 0 to 1

- The solution is to transform the response variable to a "log-odds" form
  - Using a link function $\ln(\frac{\pi}{1-\pi})$, that maps a {0, 1} variable to {-∞, +∞}
  - This model gives us the probability of an event happening (success, failure)
  - Interpreting the odds results: how many times more likely is an event going to happen
    - Probability of happening = 80%, probability of not happening = 20%
    - Odds = 80%/20% = 4 to 1

LOGISTIC REGRESSION II

- The 2nd model that is widely used in academia and industry
- The logistic regression model is set up as a linear combination of independent variables

$$\ln\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta_1 x_i + \cdots + \beta_k x_k$$

- Independent variables can be quantitative or categorical (similar to linear regression)
  - Income
  - Qualifications
  - Job Title
  - Etc…

EXERCISE

- Upload "logistics regression.ipynb" into your dashboard
- Follow the comments and codes in the notebook
- Note that for predictive projects:
  - We need to split the data into train and test set
  - If we used a train models to predict on any observations in the training set, we are going to get very good results (this is in-sample bias)
  - To get an objective accuracy test, we need to use the trained model on a set of observations that the model has not been trained on
- Load the dataset on Weekly S&P Stock Market and see if you can create a logistics model to predict the direction of the stock

QUESTIONS?

Email any queries to
jackhong@smu.edu.sg